statstutor

community project

encouraging academics to share statistics support resources

stcp-karadimitriou-scatterR

The following resources are associated:

Csv dataset 'Birthweight reduced.csv', scatterplots script file, Correlation in R, Regression in R,

Scatterplots in R

Variables: Two continuous (scale) variables.

Common Applications: Assessing the strength of a linear relationship between two continuous variables.

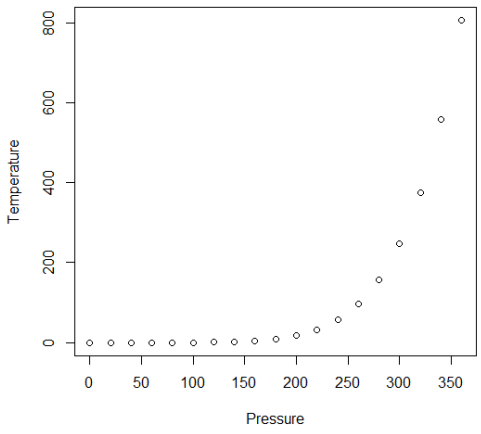
Scatterplots

Look for these things when interpreting a scatterplot:

- Is the relationship weak, moderate or strong
- Is the relationship linear?
- If yes, is it positive or negative?
- Are there any outliers?

In the plot to the left, the relationship between kilometers run per week and weight in kilometers is investigated. Generally, there is a moderate negative relationship (as weight goes down, km per week goes up) which is approximately linear. There is one outlier (red dot) but it is not extreme enough to be a data entry error.

Correlation measures the strength of a linear relationship which means the pattern looks roughly like a line. The graph to the right (air pressure data) is an example of a non-linear relationship as although there is a clear relationship, the points do not form a line.



© Sofia Maria Karadimitriou and Ellen Marshall
University of Sheffield

Reviewer: Jim Bull
University of Swansea

Data: The data set '*Birthweight reduced.csv*' contains details of 42 babies and their parents at birth. The research question is which factors affect birth weight. The dependent variable is Birth weight (lbs) and the independent variables for this sheet are gestational age of the baby at birth (in weeks) and whether or not the mother smokes (smoker). Note:You do not always need a dependent variable when looking for an association between two variables.

Steps in R

When carrying out regression, scatterplots should be produced for each independent with the dependent so see if the relationship is linear (scatter forms a rough line). Binary variables can be distinguished by different markers on scatterplots which helps to investigate patterns within groups. Open the birthweight reduced dataset which is saved as a csv file and call it `birthweightR`. You will need to change the command depending on where you have saved the file.

```
birthweightR<-read.csv("D:\\Birthweight reduced.csv",header=T)
```

Tell R we are using the birthweight dataset until further notice using attach. This means that 'Gestation' can be used instead of birthweightR$Gestation.

```
attach(birthweightR)
```

R assumes all numeric values are continuous so tell it that 'smoker' is a factor and attach labels to the categories (for example 0 in smoker means the mother is a non-smoker).

The factor command uses variable<-factor(variable,c(categorynumbers),labels=c(category names)).

```
smoker<-factor(smoker,c(0,1),labels=c('Non-smoker','Smoker'))
```

Creating a simple scatterplot

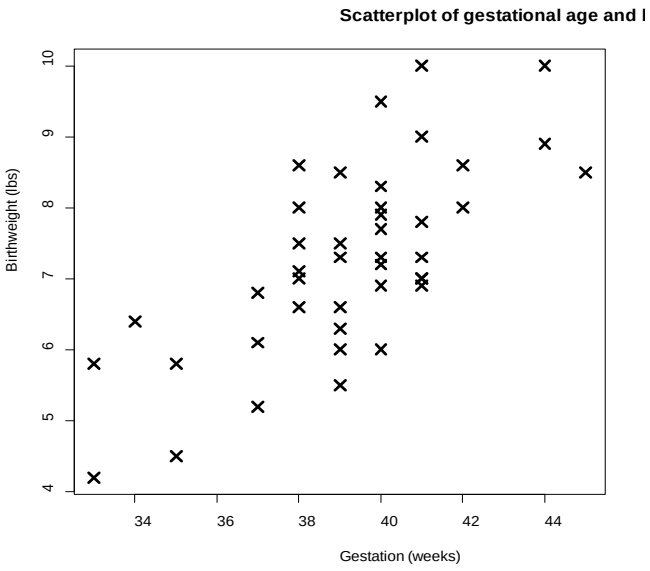
To produce a basic scatterplot showing the relationship between two scale variables use:
`plot(x variable,y variable)e.g. plot(Gestation,Birthweight)`

The title of the plot can be changed using the `main` attribute and x and y axis labels using `xlab''` and `ylab''` e.g. `plot(...,xlab='Gestation (weeks)',ylab='Birthweight(lbs)')`.

The attribute `pch` changes the shape of the scatter e.g. `pch=4` gives crosses instead of dots. `cex` magnifies the scatter e.g. `cex=2` doubles the size and `lwd` changes the thickness of the line.

```
plot(Gestation,Birthweight,
main='Scatterplot of gestational age
and birthweight',xlab='Gestation
(weeks)',ylab='Birthweight
(lbs)',pch=4,cex=1.5,lwd=3)
```

You can also alter the limits of the axes of the plot. For example, to change the x axis to show values from 30 to 50 and for the y axis from 3 to 11, add the attribute
`plot(...,xlim=c(30,50),ylim=c(3,11)).`



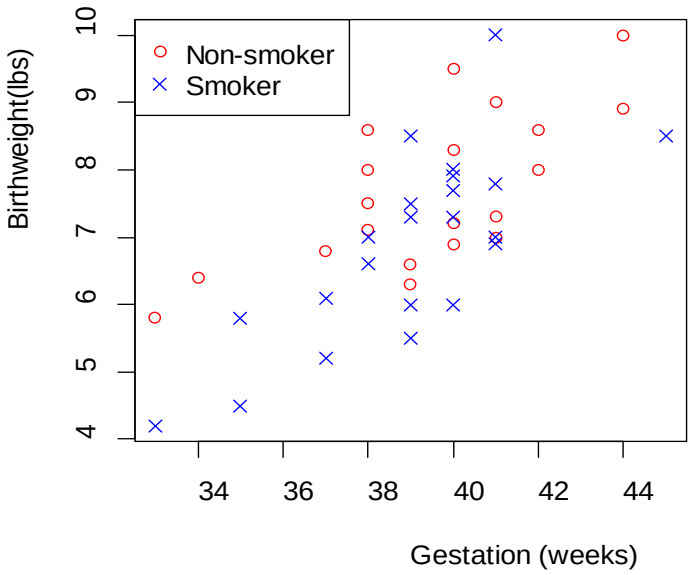
The simple scatterplot shows that there is a strong positive linear relationship between birthweight and gestational age.

Creating a scatterplot to compare groups

It is a good idea to change the color of the scatter for one group to make group comparison clearer. You can define the colors through `col` attribute in the `plot()` command and define different colours for each groups that you have. You can also 'play' with the type of dots that you get by adding the attribute `pch` will give you two forms of dots for the two different categories, for instance `pch=c(1,4)`.

```
plot(Gestation,Birthweight,col=c('red','blue')[smoker],main='Scatterplot
by smoker', pch=c(1,4)[smoker],xlab='Gestation
(weeks)',ylab='Birthweight(lbs)')
```

Scatterplot by smoker



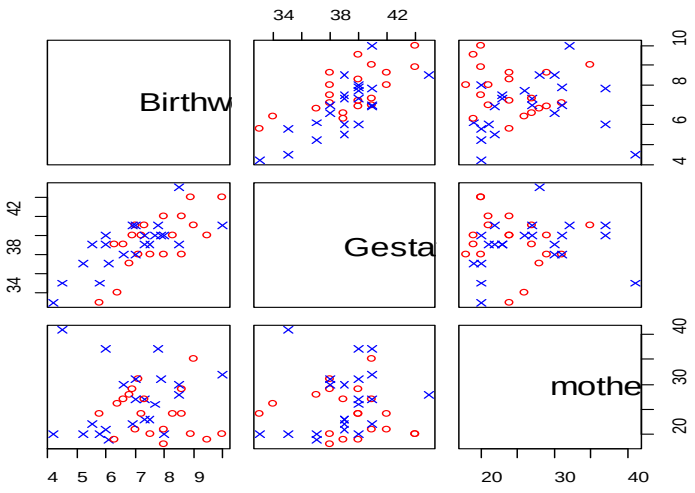
You will need to add a legend with the labels for each group. For the legend, you have a variety of choices of where to place it; type `?legend` to see the choices. As you can see, you define the legend to take the names of the categories for smokers, then you define again the colors that you put previously in your plot and finally, the shape of the scatter to be used.

```
legend(x="topleft", legend =
levels(smoker), col=c('red','blue'),
pch=c(1,4))
```

From the scatterplot, it looks like the babies of smokers tend to be lighter at each gestational age and both groups have a positive relationship between gestational age and birth weight.

It is also possible to produce a scatterplot matrix which is a table containing multiple scatterplots showing all pairwise relationships between variables. This is particularly useful for multiple regression where one of the assumptions is that each independent variable shows a linear relationship with the dependent variable. It is also important that pairs of independent variables are not strongly related.
Create a scatterplot matrix by placing all of the variables to the right of the ~ attribute.

Birth weight scatterplo



As with the simple scatterplot, different groups can be displayed with different colours using from the *col* attribute and shapes using the *pch* attribute.

```
pairs(~Birthweight+Gestation+motherage,main='Birth weight scatterplot
matrix',col=c('red','blue')[smoker],pch=c(1,4)[smoker])
```

From the scatterplot matrix, we can see that the strongest relationship is between gestation and birth weight and that smokers tend to have lighter babies. However, there is no relationship between the age of the mother and both birthweight and gestational age for all the mothers generally and there is no pattern for smoking and non smoking mothers.

Adding a line of best fit through the data

It is also possible to fit a regression line through the data using the command `abline()` to see how well a linear model could fit the data. The `lwd` attribute in the legend, specifies the thickness of the line to be displayed and `col=' '` specifies the colour of the line. The thickness of line can be specified in both the abline and lines commands, for instance, `abline(....,lwd=1)`, `lines(...,lwd=2)`.

```
plot(Gestation,Birthweight,main='Scatterplot of gestational age and
birthweight')
abline(lm(Birthweight~Gestation),col='red',lwd=2)
```

From the regression line, we can see that a linear model would fit the data very well. Simple linear regression can be carried out to see if there is a significant relationship between gestational age and birth weight.